Reinforcement Learning: An Introduction

Solutions: Chapter 4

Mrinank Sharma

March 29, 2020

Exercise 4.1

$$q_\pi(11, \texttt{down}) = -1 + v_\pi(\texttt{terminal}) = -1. \tag{1}$$
$$q_\pi(7, \texttt{down}) = -1 + v_\pi(11) = -1 - 14 = -15. \tag{2}$$

Exercise 4.2: MDP Modifications

Unchanged dynamics from current states. In this case, the value of these states from these policies is unchanged, we've simply introduced a new state. We can calculate the value of this state using the Bellman equations:

$$v_\pi(15) = -1 + \frac{1}{4}[-22 - 20 - 14 + v_\pi(15)] \Rightarrow v_\pi(15) = -20. \tag{3}$$

Dynamics changed from state 13. In general, we'd expect this to change the value of state 13, which would propagate to all of the other states, meaning that we'd have to recalculate the values. Let's pretend that we were running iterative policy evaluation using the previous value function as our initialisation. Let's update the value for state 13:

$$v_\pi(13) = -1 + \frac{1}{4}[-20 - 22 - 14 \quad \underbrace{-20}_{\text{Estimate for } v_\pi(15)} \quad] = -20 \tag{4}$$

The value of state 13 **hasn't changed!** Therefore, the value of state 15, and all of the other states also won't change. We've recalculated values without having to solve those annoying equations, which is nice.

Exercise 4.3: Iterative Action-Value Evaluation

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r | s, a)[r + \gamma \sum_{a \in \mathcal{A}(s')} \pi(a'|s')[q_\pi(s', a')]], \end{aligned} \tag{5}$$

which can straightforwardly be turned into an iterative update equation.